IBM Applied Data Science Capstone Project

ALTERNATIVE CITIES TO LIVE IN THE US FROM THE TOP 3 STATES OF POPULATION

Project Report

Abstract

Find the alternative cities to live in the United States of America from the top three states of Population based on the 2019 data.

Srinivasan Sugumar

Contents

Introduction

With the rise in popularity of the modern lifestyle globally, settling in developed nation has been a common nature for all sectors of people. So, analyzing the population of the most common cities and choosing the best possible scenario seems to be a right topic for the project.

The Data is collected from the Wikipedia about the cities of the three states. Now, the data collected is queried for the location and on the popularity of the venue for the proper classification of the cities to be chosen by the personals.

Background

Over the past decade, New York, California and Illinois are the fastest growing states in the United States. These are the most populous incorporated states of the United States including designations, including city, town, village, borough and municipality. (Based on the Wikipedia source for the year 2019 https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population)

In this project, my focus is on studying about the towns and cities of these states, collecting the data sources of the population and related details about the three states and analyzing them to find and guide a better place among these three states that have all the satisfying amenities like better workplace, living space and settling with family.

Problem

Once the settlements increase in the towns and cities of these three states, the cost of living are skyrocketing. But, due to the work opportunities, we are pushed to select the best out of the situation within these places. The cities like Los Angeles, San Diego and San Jose in California, Hempstead, Brookhaven and Islip in New York and Chicago, Aurora and Rockford in Illinois are the most populous among these states. Through this project, I am trying to give a suggestion for people who would love to settle in these states other than these most populous cities but still avail all or the utmost features of the above cities of these three popular states of USA.

Audience

The primary audience of this study might include realtors and potential home buyers/renters in these regions. The findings could also be used by the entrepreneurs looking to open new businesses or even a way of fostering outreach and partnerships among the municipal chambers of commerce.

Data

I. Sources

To obtain a list of cities of these states, we'll scrape Wikipedia for a list of cities in New York, California and Illinois. We'll use the Foursquare venue recommendation to obtain a list of the most popular venues for each city and query location data (latitude/longitude) using the Mapquest Geocoding in order to map all the cities and visualize the clusters. Check out the References for the data sources used in the project.

II. Data Cleansing

From the data collected in Wikipedia, we need only the city names of the three respective states and so the rest of the data are not taken into account and the three respective state names are added.

Preparation

The data collected from the Wikipedia pages are stored as Pandas Data Frame for easier use and analysis. The Data Frame is used to store the city, state, longitude and latitude values and combine the three states dataframe to a single dataframe representing all the cities taken into account.

Then, using Mapquest API, the venues are found, can be termed as geocoding.

Finally, using the Folium mapping library, the final set of data(cities) are made ready for further analysis and process.

The following pandas dataframe shows the top 3 US States in Population as of 2019 (estimated)

	2019rank	City	State[c]	2019estimate	2010Census	Change	2016 land area	2016 land area.1	2016 population density	2016 population density.1
0	1	New York[d]	New York	8336817	8175133	+1.98%	301.5 sq mi	780.9 km2	28,317/sq mi	10,933/km2
1	2	Los Angeles	California	3979576	3792621	+4.93%	468.7 sq mi	1,213.9 km2	8,484/sq mi	3,276/km2
2	3	Chicago	Illinois	2693976	2695598	−0.06%	227.3 sq mi	588.7 km2	11,900/sq mi	4,600/km2
3	4	Houston[3]	Texas	2320268	2100263	+10.48%	637.5 sq mi	1,651.1 km2	3,613/sq mi	1,395/km2
4	5	Phoenix	Arizona	1680992	1445632	+16.28%	517.6 sq mi	1,340.6 km2	3,120/sq mi	1,200/km2

We will collect each state's towns and cities one by one and consolidate into a single pandas dataframe for easiness and analysis

All the three states towns and cities data are consolidated into a

Single pandas dataframe as a first step.

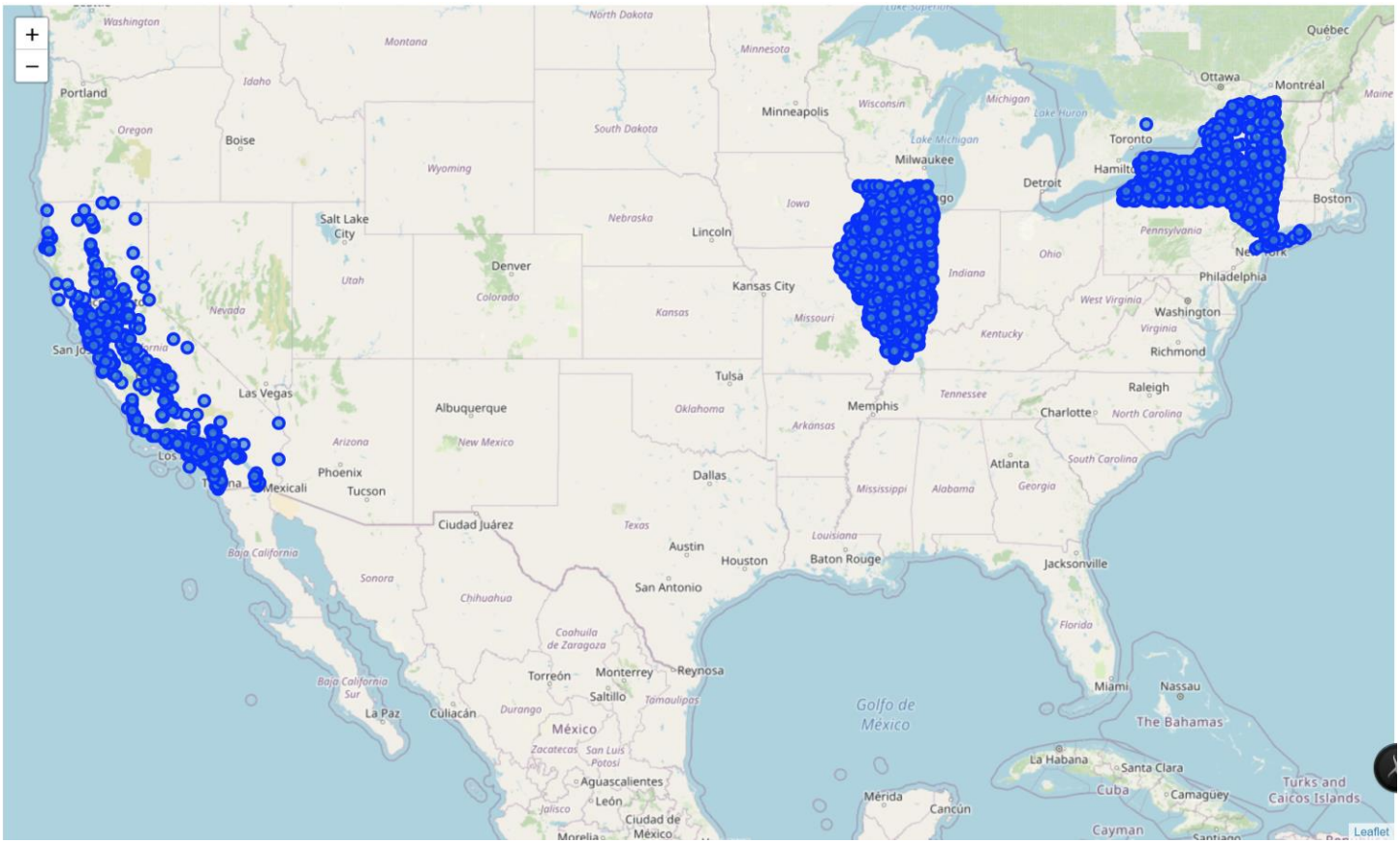
	City	State
0	Adams	NY
1	Addison	NY
2	Afton	NY
3	Alabama	NY
4	Albion	NY
...	...	...
2709	Yale	IL
2710	Yates City	IL
2711	Yorkville†	IL
2712	Zeigler	IL
2713	Zion	IL

Then the Mapquest API is used to find the latitude and longitude of the respective towns and cities of the three states in order to complete the geocoding (Tagging) process.

	City	State	Latitude	Longitude
0	Adams	NY	43.8105	-76.0235
1	Addison	NY	42.1051	-77.2343
2	Afton	NY	42.2295	-75.5245
3	Alabama	NY	43.0964	-78.391
4	Albion	NY	43.2463	-78.1938
...	...	...	...	...
2709	Yale	IL	39.121	-88.0249
2710	Yates City	IL	40.7787	-90.0146
2711	Yorkville†	IL	41.6414	-88.4469
2712	Zeigler	IL	37.8967	-89.0554
2713	Zion	IL	42.4571	-87.8253

2714 rows × 4 columns

Once the latitude and longitude geocoding process are done then the folium package are used to plot the data into a map and the same is shown below



Methodology

We have collected the data that we need to analyze.

We will use unsupervised machine learning algorithm called k-means clustering that enables us to partition observations into a specified number of clusters in order to discover underlying patterns.

With the data, we will find the top 5 venue categories for each city (based on occurrences in the dataset), and use that as each city's vector profile for finding similarities with other cities.

First, we need to calculate the average frequency for each venue category across each city. We can quickly do this with a Pandas dataframe by converting each venue category into a Boolean (yes/no) column.

The following illustration shows how the Mapquest and Foursquare APIs are used to get the venue details for all the towns and cities of top 3 states namely New York, California and Illinois.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Adams	43.810509	-76.023503	Gram's Diner	43.808869	-76.024474	Diner
1	Adams	43.810509	-76.023503	Hometown Pizzeria	43.809169	-76.024676	Pizza Place
2	Adams	43.810509	-76.023503	Jreck Subs	43.809642	-76.021185	Sandwich Place
3	Adams	43.810509	-76.023503	Mercers	43.809660	-76.021007	Convenience Store
4	Adams	43.810509	-76.023503	Dp Bartlett & Sons	43.814521	-76.023847	Construction & Landscaping
...	...	...	...	...	...	...	...
24361	Zion	42.457095	-87.825306	Dog House	42.454251	-87.825905	Hot Dog Joint
24362	Zion	42.457095	-87.825306	Star Lite Restaurant	42.453511	-87.825190	Restaurant
24363	Zion	42.457095	-87.825306	Starlite	42.453465	-87.825065	Breakfast Spot
24364	Zion	42.457095	-87.825306	Starlight Country Restaurant	42.453400	-87.825200	Restaurant
24365	Zion	42.457095	-87.825306	Randy Nebel - State Farm Insurance Agent	42.453063	-87.825258	Insurance Office

24366 rows × 7 columns

Each town and city have a number of venues that are useful but when we want to narrow down our analysis, we need only the relevant and the most wanted venues, refer the below for the entire venue lists from the towns and cities of the three states.

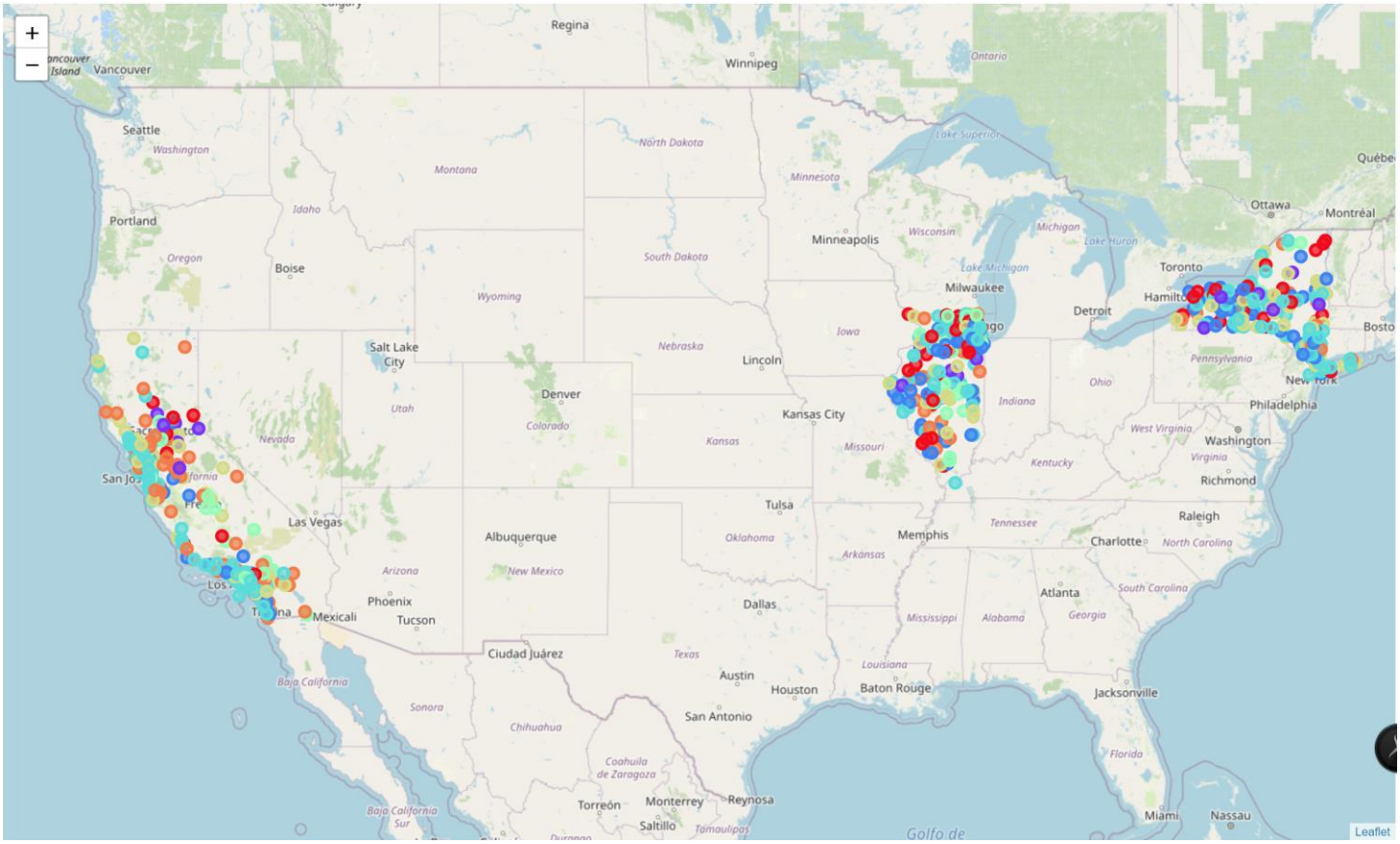
	City	State	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Addison	NY	42.1051	-77.2343	2.0	Pizza Place	Ice Cream Shop	Italian Restaurant	Thai Restaurant	New American Restaurant
4	Albion	NY	43.2463	-78.1938	2.0	Pizza Place	Post Office	Sandwich Place	Donut Shop	Liquor Store
5	Albion	NY	43.2463	-78.1938	2.0	Pizza Place	Post Office	Sandwich Place	Donut Shop	Liquor Store
8	Alexandria	NY	44.3368	-75.919	5.0	Resort	American Restaurant	Pizza Place	Steakhouse	Sporting Goods Shop
9	Alfred	NY	42.2547	-77.7904	0.0	Bar	Café	Movie Theater	Grocery Store	Pharmacy
17	Amherst	NY	42.9791	-78.7993	3.0	Furniture / Home Store	Rental Car Location	Video Store	Dessert Shop	Rental Service
27	Arcadia	NY	43.0466	-77.0953	2.0	Racetrack	Hotel	Food Truck	Pizza Place	Fountain
34	Athens	NY	42.2601	-73.8089	0.0	Bar	American Restaurant	Harbor / Marina	Insurance Office	Brewery
39	Aurora	NY	42.754	-76.7024	6.0	Mexican Restaurant	Bed & Breakfast	Hotel	Tattoo Parlor	Flea Market
43	Avon	NY	42.9127	-77.7459	0.0	Bar	Gas Station	American Restaurant	Bakery	Gym
44	Babylon	NY	40.6958	-73.3257	3.0	American Restaurant	Spa	Mexican Restaurant	Italian Restaurant	Asian Restaurant
46	Baldwin	NY	40.663	-73.6107	2.0	Pharmacy	Bar	Movie Theater	Chinese Restaurant	Lounge
51	Barrington	NY	42.5503	-77.0564	2.0	Pizza Place	Sandwich Place	Furniture / Home Store	Donut Shop	Video Store
53	Batavia	NY	42.9969	-78.186	3.0	Coffee Shop	Bar	Taco Place	Fast Food Restaurant	American Restaurant

Based on the current population we decided to filter more than 10 venues from each towns and cities and top 5 venue categories for each towns and cities in order to filter the cities for the most suited.

	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	count
5	Addison	42.105085	-77.234330	Dollar General	42.103854	-77.235717	Discount Store	29
6	Addison	42.105085	-77.234330	7-Eleven	42.105657	-77.233682	Convenience Store	29
7	Addison	42.105085	-77.234330	Main Grill	42.106368	-77.234366	Bar	29
8	Addison	42.105085	-77.234330	Sugar Creek	42.105698	-77.233795	Gas Station	29
9	Addison	42.105085	-77.234330	China Wok	42.106501	-77.233801	Chinese Restaurant	29
...	...	...	...	...	...	...	...	...
24361	Zion	42.457095	-87.825306	Dog House	42.454251	-87.825905	Hot Dog Joint	26
24362	Zion	42.457095	-87.825306	Star Lite Restaurant	42.453511	-87.825190	Restaurant	26
24363	Zion	42.457095	-87.825306	Starlite	42.453465	-87.825065	Breakfast Spot	26
24364	Zion	42.457095	-87.825306	Starlight Country Restaurant	42.453400	-87.825200	Restaurant	26
24365	Zion	42.457095	-87.825306	Randy Nebel - State Farm Insurance Agent	42.453063	-87.825258	Insurance Office	26

18127 rows × 8 columns

The textual data looks fine when we have limited data but when the data is bulk and we need to understand how and where the locations are placed and what the neighborhood places are, a graphical representation will be more appropriate. Let's plot the venue details from the towns and cities into map for more meaningful data. Refer the below map for the entire data.



We are using K-Means clustering unsupervised algorithm for the classification of data (Venues) with the cluster centroid 7 for meaningful data classification. These clustering data are furnished as below

Cluster 0:

This cluster consists 65 outlier towns and cities from the three states. We can see from the venue data and most of primary venues are Bar.

Cluster 0: Outliers

This cluster consists 65 outlier towns and cities from the three states. We can see from the venue data and and most of primary venues are Bar.

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', 1)

cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 0, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
9	Alfred	NY	Bar	Café	Movie Theater	Grocery Store	Pharmacy
34	Athens	NY	Bar	American Restaurant	Harbor / Marina	Insurance Office	Brewery
43	Avon	NY	Bar	Gas Station	American Restaurant	Bakery	Gym
71	Binghamton	NY	Bar	Café	Coffee Shop	Diner	New American Restaurant
105	Byron	NY	Convenience Store	Sandwich Place	Bar	Grocery Store	American Restaurant
146	Chatham	NY	Bar	Fast Food Restaurant	Sandwich Place	Insurance Office	Gas Station
172	Clinton	NY	Bar	Sandwich Place	Pizza Place	Pharmacy	Boutique
173	Clinton	NY	Bar	Sandwich Place	Pizza Place	Pharmacy	Boutique
318	Geneva	NY	Bar	Pizza Place	Gift Shop	Coffee Shop	Hotel
417	Islip	NY	Bar	Food Court	Diner	Sandwich Place	Movie Theater

Cluster 1:

Cluster 1 classified by hotels/resorts, restaurants and nightlife (bars, breweries, etc). Many of the cities on this list are vacation destinations and/or popular weekend getaway spots.

Cluster 1: Vacation destinations

Cluster 1 classified by hotels/resorts, restaurants and nightlife (bars, breweries, etc). Many of the cities on this list are vacation destinations and/or popular weekend getaway spots.

```
pd.set_option('display.max_rows', None)
pd.set_option('display.max_columns', None)
pd.set_option('display.width', None)
pd.set_option('display.max_colwidth', 1)

cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 1, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
93	Brookfield	NY	Park	Irish Pub	Pizza Place	Sports Bar	Business Service
317	Geneseo	NY	Park	Bar	American Restaurant	Pizza Place	Café
413	Inlet	NY	Park	Coffee Shop	Grocery Store	Bookstore	Ice Cream Shop
488	Lyons	NY	Bar	Park	Sandwich Place	Fast Food Restaurant	Supermarket
606	Olean	NY	Steakhouse	Bar	Convenience Store	Café	Mexican Restaurant
674	Plymouth	NY	Post Office	Convenience Store	New American Restaurant	Hotel	Diner
736	Salina	NY	Park	Coffee Shop	American Restaurant	Supermarket	Thai Restaurant
875	Waterford	NY	Park	Café	American Restaurant	Chinese Restaurant	Sandwich Place
904	Wheatland	NY	Park	Bar	Bridge	River	Café

Cluster 2:

Pizza place is common among nearly all of the cities in this cluster. It looks like Pizza Place, Mexican Restaurant, Chinese Restaurant, American Restaurant, and Bar are all grouping together here. The towns and cities on this list tend to be larger than bedroom communities, but somewhat smaller than major urban centers.

Cluster 2: : Restaurant towns and cities

Pizza place is common among nearly all of the cities in this cluster. It looks like Pizza Place, Mexican Restaurant, Chinese Restaurant, American Restaurant, and Bar are all grouping together here. The towns and cities on this list tend to be larger than bedroom communities, but somewhat smaller than major urban centers.

```
# Cluster 2
cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 2, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
1	Addison	NY	Pizza Place	Ice Cream Shop	Italian Restaurant	Thai Restaurant	New American Restaurant
4	Albion	NY	Pizza Place	Post Office	Sandwich Place	Donut Shop	Liquor Store
5	Albion	NY	Pizza Place	Post Office	Sandwich Place	Donut Shop	Liquor Store
27	Arcadia	NY	Racetrack	Hotel	Food Truck	Pizza Place	Fountain
46	Baldwin	NY	Pharmacy	Bar	Movie Theater	Chinese Restaurant	Lounge
51	Barrington	NY	Pizza Place	Sandwich Place	Furniture / Home Store	Donut Shop	Video Store
89	Brighton	NY	Pizza Place	Optical Shop	Dance Studio	Kids Store	Pharmacy
90	Brighton	NY	Pizza Place	Optical Shop	Dance Studio	Kids Store	Pharmacy

Cluster 3:

In addition to coffee shops being the prevalent venue type, the cities in this cluster are characterized by a diverse set of amenities, indicative of larger urban centers.

Cluster 3: Coffee shop towns and cities

In addition to coffee shops being the prevalent venue type, the cities in this cluster are charaterized by a diverse set of ammenities, indicative of larger urban centers.

```
# Cluster 3
cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 3, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
17	Amherst	NY	Furniture / Home Store	Rental Car Location	Video Store	Dessert Shop	Rental Service
44	Babylon	NY	American Restaurant	Spa	Mexican Restaurant	Italian Restaurant	Asian Restaurant
53	Batavia	NY	Coffee Shop	Bar	Taco Place	Fast Food Restaurant	American Restaurant
114	Campbell	NY	Mexican Restaurant	Park	Ice Cream Shop	Breakfast Spot	Italian Restaurant
138	Cazenovia	NY	Café	Diner	Pizza Place	Thai Restaurant	Bank
149	Cheektowaga	NY	Intersection	Mobile Phone Shop	Donut Shop	Fried Chicken Joint	Car Wash
157	Chili	NY	Flower Shop	Coffee Shop	Pharmacy	Community Center	Video Store
168	Clayton	NY	Harbor / Marina	Park	Sandwich Place	Café	Pizza Place
170	Clifton	NY	Bus Stop	Grocery Store	Deli / Bodega	Pool	Food
181	Coldspring	NY	Café	Gift Shop	Antique Shop	Grocery Store	Pet Store
198	Corning	NY	Café	American Restaurant	Mexican Restaurant	Hotel	Historic Site

Cluster 4:

The unifying characteristic among these cities is that Fast Food Restaurant is the prominent venue type. These are smaller cities and bedroom communities that tend to be located between larger cities with more amenities. At first glance, a "Fast food towns and cities" might not seem particularly attractive, but this cluster deserves further exploration for prospective home buyers looking for more seclusion and lower real estate prices.

Cluster 4: Fast food towns and cities

The unifying characteristic among these cities is that Fast Food Restaurant is the prominent venue type. These are smaller cities and bedroom communities that tend to be located between larger cities with more ammentiies. At first glance, a "Fast food towns and cities" might not seem particularly attractive, but this cluster deserves further exploration for prospective home buyers looking for more seclusion and lower real estate prices.

```
# Cluster 4
cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 4, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
158	Cicero	NY	Fast Food Restaurant	Mexican Restaurant	Pizza Place	Train Station	Donut Shop
185	Colton	NY	Pizza Place	Breakfast Spot	Fast Food Restaurant	Mexican Restaurant	Sushi Restaurant
306	Fulton	NY	Fast Food Restaurant	Deli / Bodega	Donut Shop	Mobile Phone Shop	Mexican Restaurant
468	Lincoln	NY	Sandwich Place	Breakfast Spot	Mexican Restaurant	Bar	Plaza
742	Santa Clara	NY	Fast Food Restaurant	Indian Restaurant	Fried Chicken Joint	Chinese Restaurant	Mexican Restaurant
946	Apple Valley	CA	Park	Fast Food Restaurant	Chinese Restaurant	Mobile Phone Shop	Pharmacy
951	Arvin	CA	Fast Food Restaurant	Pizza Place	Chinese Restaurant	Sandwich Place	Supermarket

Cluster 5:

The yield for the final cluster was a couple more restaurants towns and cities, more are in New York state.

Cluster 5: More Restaurants

The yield for the final cluster was a couple more restaurants towns and cities, more are in New York state.

```
# Cluster 5
cites_venus_data.loc[cites_venus_data['Cluster Labels'] == 5, cites_venus_data.columns[[0] + [1] + list(range(5, cites_venus_data.shape[1]))]]
```

	City	State	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
8	Alexandria	NY	Resort	American Restaurant	Pizza Place	Steakhouse	Sporting Goods Shop
55	Bedford	NY	Deli / Bodega	Café	Concert Hall	American Restaurant	Flower Shop
118	Canandaigua	NY	Pub	Brewery	American Restaurant	Italian Restaurant	Business Service
136	Catskill	NY	Café	Diner	American Restaurant	Ice Cream Shop	Mexican Restaurant
154	Chester	NY	Bar	Italian Restaurant	Brewery	Trail	American Restaurant
155	Chester	NY	Bar	Italian Restaurant	Brewery	Trail	American Restaurant
226	Delaware	NY	American Restaurant	Bar	Vietnamese Restaurant	Park	Bed & Breakfast

Conclusion

From the three states, namely New York, California and Illinois, we have a total of 2714 cities, out of which only 644 cities are listed with the Foursquare venue data. A Foursquare query of venues in these cities yielded 24,366 venues, however, we need to filter out cities with fewer than 10 venues, as their profile later proved insufficient for meaningful clustering. After filtering out those cities, only 644 remained.

The 644 cities used in the final analysis represented 18127 venues and 462 unique venue types. We used the k-means clustering algorithm to group them into six distinct clusters, however only four of those clusters were truly meaningful in terms of revealing insights among our dataset that we could use to answer the original question of our problem search.

How can the residents of these states identify similar cities as prospective places to move? The results of our analysis certainly provide an idea for these residents who are planning for moving to other cities with similar life benefits.

Throughout the process of this study, we uncovered limitations in comprehensively addressing the problem at hand. We also found interesting patterns among the refined dataset of larger cities with an adequate amount of Foursquare venue data.

Next step in the process might be to supplement the data to cluster cities with additional sources like average home price, population size etc. so as to retain and cluster a full list of the cities with finer - grained grouping patterns for cities with ample Foursquare venue data.

Recommendations

It is clearly known that people, investors and real estate persons can focus on these towns and cities. We found that these cities are having all the needed facilities with the available data from the Foursquare API. We would like to have more data and analysis to fine tune the model for better results.

References

[1.] https://en.wikipedia.org/wiki/List_of_United_States_cities_by_population

[2.] https://en.wikipedia.org/wiki/List_of_towns_in_New_York

[3.] https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_California

[4.] https://en.wikipedia.org/wiki/List_of_municipalities_in_Illinois

[5.] https://api.foursquare.com/v2/venues/explore?

[6.] https://www.mapquestapi.com/geocoding/v1/address?